A User Interface-Level Integration Method for Multiple Automatic Speech Translation Systems

Seiya Osada[1], Kiyoshi Yamabana[1], Ken Hanazawa[1], Akitoshi Okumura[1]

[1] Media and Information Research Laboratories
NEC Corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, Japan
{s-osada@cd, k-yamabana@ct, k-hanazawa@cq, a-okumura@bx}.jp.nec.com

Abstract. We propose a new method to integrate multiple speech translation systems based on user interface-level integration. Users can select the result of free-sentence speech translation or that of registered sentence translation without being conscious of the configuration of the automatic speech translation system. We implemented this method on a portable device.

Keywords: speech translation, user interface, machine translation, registered sentence retrieval, speech recognition

1 Introduction

There have been many researches on speech-to-speech translation systems, such as NEC speech translation system[1], ATR-MATRIX[2] and Verbmobil[3]. These speech-to-speech translation systems include at least three components: speech recognition, machine translation, and speech synthesis. However, in practice, each component does not always output the correct result for various inputs.

In actual use of a speech-to-speech translation system with a display, the speaker using the system can examine the result of speech recognition on the display. Accordingly, when the recognition result is inappropriate, the speaker can correct errors by speaking again to the system. Similarly, when the result of speech synthesis is not correct, the listener using the system can examine the source sentence of speech synthesis on the display.

On the other hand, the feature of machine translation is different from that of speech recognition or speech synthesis, because neither the speaker nor the listener using the system can confirm the result of machine translation. Thus, an error in the machine translation is critical in a speech-to-speech translation system.

Instead of machine translation, there is a parallel text-based translation which uses parallel bilingual sentences registered in the system. This retrieves the corresponding translation by referring to the registered source sentence. However, in parallel text-based translation, although the quality of translation is largely guaranteed, only a limited number of sentences can be translated, because it is impossible to cover all the utterances of users by the parallel text of source language. When no registered sentences correspond to what the user says, he or she has to choose a most preferable one among retrieved sentences that roughly reflect the utterance; otherwise the speaker may give up the attempt.

Accordingly, an integrated method which is easier to use is required in which the accuracy of translation becomes compatible with coverage by integrating these two translation components. This paper proposes a new method to unify the automatic speech translation system with free-style sentence translation component and that with parallel text-based translation component at the user interface-level in order to further ease the operation of users compared to the previous unified approach.

2 Previous approach: Method-level integration

To solve these problems on the machine translation and the parallel text-based translation, there is a system which integrates two components -- for free-sentence translation and for parallel text-based translation [4].

This system (see Figure 1) has a speech recognition component, a speech synthesis component and two translation components – a free-sentence translation component and a parallel text-based translation component. The free-sentence translation component provides conventional machine translation. The parallel text-based translation component has two components for registered sentence retrieval and registered sentence translation. Through the registered sentence retrieval, corresponding sentences are retrieved from the source language corpora. Subsequently, through the registered sentence translation, the corresponding registered translations are chosen from the parallel corpora.

As shown in Figure 1, this system requires the user to operate as many as three times from the time when he or she utters to this machine to the time when the partner hears the result of translation. After speaking to the machine in the first operation, the user is required to determine whether the result of speech recognition should be translated by free-style sentence translation or by parallel text-based translation in the second operation. The third operation refers to the cases when parallel text-based translation is selected in the second operation. The user needs to determine whether to choose one from among the results of parallel text-base translation. If the utterance does not hit any registered sentences, the user can choose free-style sentence machine translation.

In regards to the second operation, one problem cannot be ignored; it may be rather difficult for users to select the way of translation, free-style sentence translation or parallel text-based translation, in the absence of decisive factors.

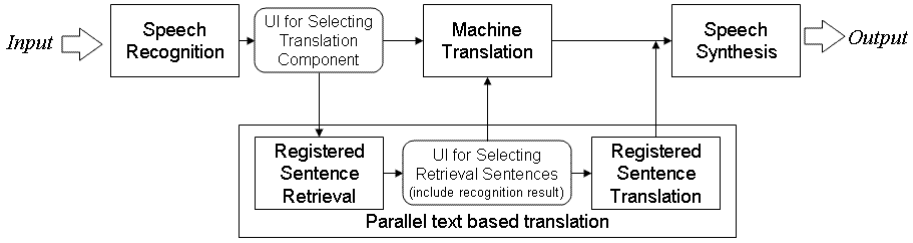


Fig. 1. Configuration of previous system

It may be rather difficult for users to understand the behaviour of this system, since the integration is based on the method-level integration in EAI[7]. The system is configured by combining component modules. More than one user interfaces are placed between these components. Many users, who expect to obtain the translation result as in the case of free-sentence translation module, may be confused about what the translation system processes.

In this system, the users should understand the configuration of Figure 1 and need to know which state the system is currently in. Additionally, the users should know to which state they want the system to proceed next.

It is desirable that a speech translation system should be executed on the portable devices such as Personal Digital Assistant (PDA). However, it is difficult to execute this system on portable devices, because this system requires a big display and a keyboard for these operations.

3 Proposed method: User interface-level Integration

From the above findings, we propose a new method in which several speech translations are integrated at the user interface-level[7] instead of the method-level as in the previous system.

We hereby show this new method consists of two automatic speech translations. One speech translation has the three components of speech recognition, machine translation and speech synthesis while the other has speech registered sentence retrieval, registered sentence translation and speech

synthesis. These two translation systems are integrated by a user interface after translation components (Figure 2).

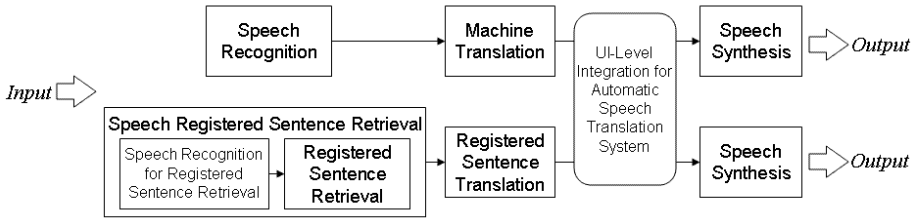


Fig. 2. Configuration of proposed method

As shown in Figure 3, the result of free-sentence translation can be listed as well as that of registered sentence translation.

The shaded portion is a free-sentence translation result of the recognized sentence. And, the rest portions are the results of registered sentence translation. The bold frame which users can move by pushing up/down keys shows a current selection.

Concerning this user interface, we must pay attention to the fact that the user interface-level integration is different from the method level integration.

In this proposed method, the translated sentence in the shaded portion may not be always correct. On the other hand, other choices guarantee the correctness of translation but the source language text may not correspond to what the user said. However, as long as users recognize this, they need not understand the configuration of the speech translation system. They are merely required to select the most suitable sentence by examining the source language text.

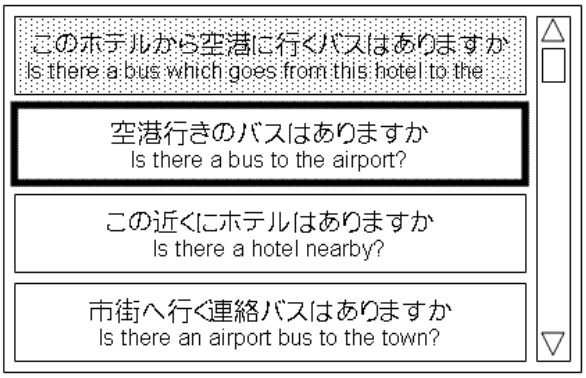


Fig. 3. User Interface-Level Integration for Automatic Speech Translation Systems
The shaded portion shows the result of speech recognition and its machine translation: "*Kono Hoteru kara kuko ni iku basu ha arimasuka*" (Is there a bus which goes from this hotel to the airport?). The following white boxes display the results of speech recognition and registered sentence retrieval: "*Kuko iki no basu ha arimasuka*" (Is there a bus to the airport?), "*Kono chikaku ni hoteru ha arimasuka*" (Is there a hotel nearby?), and "*shigai he iku renraku basu ha arimasuka*" (Is there an airport bus to the town?).

In other words, the user can select a translated sentence only by reading the source language text as long as he or she recognizes that the accuracy of translation is not always 100% on the shaded portion of the display. Also, the user doesn't have to be conscious which translation module worked on each translated sentence. Accordingly, operations should become easier on the speech-to-speech translation system.

Furthermore, this user interface is compact and can be operated in up/down keys and a few keys as in figure 3.

4 Implementation on a portable device

We have implemented the proposed method on the portable device. The hardware specifications of the portable device, the configuration of system and the outline of each component are shown below:

4.1 Hardware

Display: 2.7inch(320×240), No touch panel
Keys: up/down, right/left, OK and Back

4.2 Configuration of Implementation

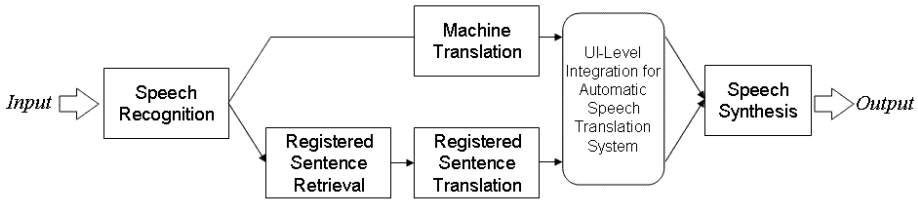


Fig. 4. Configuration of Implementation

As shown in Figure 4, our proposed method is composed of speech recognition, machine translation, registered sentence retrieval/translation and speech synthesis. Though the proposed method has two speech recognitions and two speech syntheses as shown in Figure 3, these recognitions and syntheses which have similar functions respectively are integrated respectively in this implementation. The result of machine translation and that of registered sentence are integrated by a user interface shown in Figure 3.

4.3 Speech recognition

The speech recognition component carries out a large vocabulary continuous speech recognition for both Japanese and English on travel conversation. The speech recognition method is based upon the acoustic model expressed by Hidden Markov Model (HMM) and the statistical language model[5].

Speech recognition accuracy was evaluated offline by wav file input. For Japanese, a total of 1000 utterances of travel conversation by 5 male speakers were used. For English, a total of 5760 utterances of travel conversation by 32 male speakers were used. The word accuracy was 96% for Japanese and 92% for English. And sentence correct was 81% for Japanese and 76% for English.

4.4 Machine translation

A common component is used for Japanese to English and English to Japanese machine translation. The direction of translation can be switched by changing the translation languages knowledge base written in the Lexicalized-Tree-AutoMata-based Grammar[6]. In the machine translation component, firstly a recognition result received from speech recognition as an input is parsed based upon the algorism enhancing the bottom-up chart method. A translated sentence is composed based upon the results of analysis as top-down.

The accuracy of translation was subjectively evaluated by two evaluators for 500 sentences randomly sampled from travel conversation corpora. The results of translation were classified into three levels;

1) “good” when the translated sentence is syntactically correct and the source language is fairly translated into the target language; 2) “understandable” when the source language is fairly translated into the target language although the level doesn’t reach “good”; 3) “bad” when the translated sentence is not understandable or can be misunderstood. Results indicated that the meaning of the source language can be presumed correctly (“good” or “understandable”) from the translated sentence in 87% of Japanese to English translations and in 93% of English to Japanese translations. And this can translate most of travel sentences within one second.

4.5 Registered sentence retrieval / translation

In a registered sentence retrieval/translation component, firstly key words are chosen from a recognition result received from speech recognition as an input. After retrieving registered sentences containing the key words, the sentence is translated into the target language. Ranking is done mainly depending on the number of key words contained. Our system contains around 7000 registered sentences in parallel corpora.

4.6 Speech synthesis

A waveform segment concatenation-based synthesis method, in which a synthesis unit is enhanced by using a large-sized database, is employed for Japanese speech synthesis. On the other hand, an over the counter component is utilized for English speech synthesis.

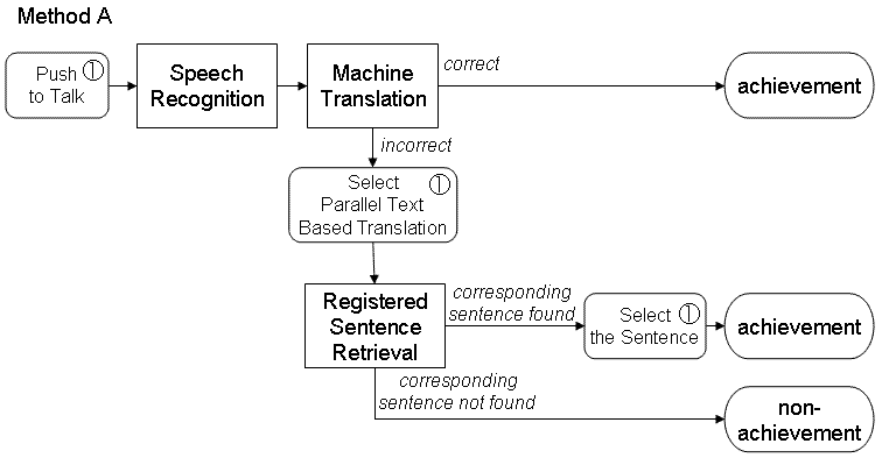
5 Evaluation of User Interface

Theoretical evaluation is made to compare the number of operations required of users between a proposed method and two other methods which have been simplified from the previous approach.

Concerning method A, the first simplified version of the previous approach, machine translation is implemented after speech recognition. If the result of machine translation is correct, no further procedure is done, however, if the result is inappropriate, registered sentence retrieval is selected.

With method B, the second simplified version of the previous approach, registered sentences are automatically retrieved after speech recognition. If the speech of user corresponds to the result of registered sentence retrieval, all the operation is over, however, if not, machine translation follows.

Above-mentioned procedure is shown by Figure 5. Users need to operate the system for boxes marked “①”.



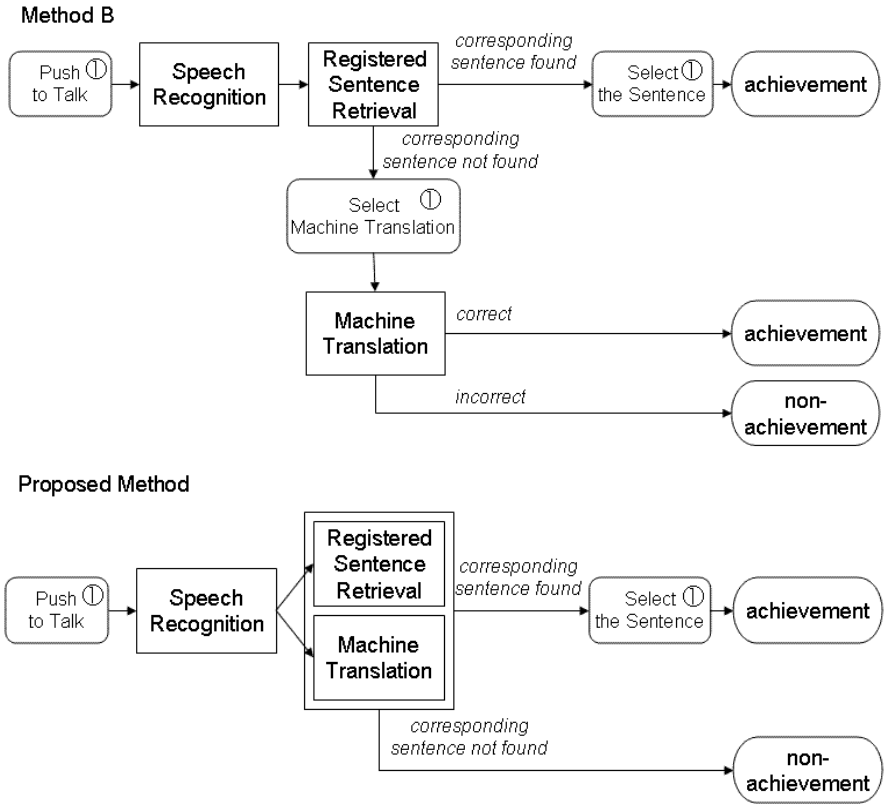


Fig. 5. Procedure on three methods

Under these figures, the number of operation required of user is specified below for the three methods.

Method A:
Once: when the result of machine translation is correct.
Three times: when the result of machine translation is incorrect and the speech of user corresponds to the result of registered sentence retrieval.
Twice: when the speech of user hits neither the result of machine translation nor that of registered sentence retrieval.

Method B:
Twice: when the result of registered sentence retrieval corresponds to the speech of user; when the result of registered sentence retrieval does not match the speech of user and the result of machine translation is appropriate; when neither the result of registered sentence retrieval nor that of machine translation reflects the speech of user.

Proposed Method:
Twice: when the speech of user corresponds to the result of machine translation or that of registered sentence retrieval.
Once: when neither the result of registered sentence retrieval nor that of machine translation represents the speech of user.

The cost, which should vary in each operation depending on the method, is taken as 1 for all the operations.

Above findings clearly show that proposed method requires fewer operations compared to method B.

The following formulae generate the average number of operation for method A and proposed method:

Method A:

$$1\ xy + 3\{\ x(1 - y)\ z + (1 - x)\ w\ \} + 2\{\ x(1 - y)\ (1 - z) + (1 - x)\ (1 - w)\}$$

Proposed Method:

$$2\ xy + 2\{\ x(1 - y)\ z + (1 - x)\ w\ \} + 1\{x(1 - y)\ (1 - z) + (1 - x)\ (1 - w)\}$$

x: the rate of sentence correct speech recognition
y: the rate of correct machine translation
z: the rate of registered sentence retrieval when the result of speech recognition is correct
w: the rate of registered sentence retrieval when the result of speech recognition is incorrect.
(The rate of machine translation is expressed as 0% for incorrect speech recognition results.)

The following inequalities are seen when the average number of operation is fewer in method A than in proposed method:

$$xy > 0.5$$

In short, the average number of operation is fewer in method A than in proposed method when xy is 50% or more, regardless of z or w. If 50% or less, proposed method requires fewer operations compared to method A.

Supposing that the rate of speech recognition and that of machine translation for method A are approximately 70% each, $xy = 0.49$ is observed. This shows that the average number of operation becomes fewer in the proposed method than in method A.

As to prototype system, the rate of correct speech recognition and that of correct machine translation are approximately 80% each for the Japanese language, resulting in xy = around 0.64 (in this system, sentence correct of speech recognition is 81% shown 4.3 and correct of machine translation is 87% shown 4.4). Accordingly, the average number of operation is fewer in method A than in proposed method.

However, for a fact, in method A users need to judge whether the result of machine translation is correct, and few users of speech translation systems are thought to make good judgments. Thus, in method A, higher is the cost that the users judge whether the result of machine translation is correct (Select Parallel Text-Based Translation in Method A shown Figure 5). In consideration of such circumstances, even though method A requires fewer operations, we think the proposed method is better.

6 Discussion

In the proposed method, the user interface is placed after the translation modules. On the other hand, the interface can also be placed before the translation modules, as a variation to the proposed method.

In the original configuration, the translation results shown on the display will help the user to choose the most suitable one, if the user has some knowledge on the target language. This situation is highly plausible for Japanese users with a Japanese to English translation system. In addition, showing the translation result along with the source text may be useful for language learning purpose.

The second configuration will be suitable if the user has no knowledge of the target language at all. In this case, only the sentence selected by the user will be translated, reducing the whole amount of processing compared to the original configuration.

7 Conclusion

We proposed a user interface-level integration method for multiple automatic speech translation systems on a portable device: 1) a free-sentence automatic speech translation system to process through speech recognition, machine translation and speech synthesis; 2) a parallel text-based translation system to obtain the target language by retrieving a parallel corpus. This new method not only has resolved problems related to a single free-sentence automatic speech translation or a single parallel text-based translation but enables users to handle without being conscious of the whole configuration of an automatic speech translation system. Moreover, the average number of operation is fewer in the proposed method than in previous approach for users of speech translation system.

Since this system is now available in one translation direction, only from Japanese to English, we are planning to further develop the system so that it can function from English to Japanese, too. Also, we would like to enhance its use for other languages in the near future.

References

1. Takao Watanabe, Akitoshi Okumura, Shinsuke Sakai, Kiyoshi Yamabana, Shinichi Doi, and Ken Hanazawa. 2000. An automatic interpretation system for travel conversation. In *Proceedings of ICSLP 2000*.
2. Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo and Seiichi Yamamoto. 1998. A Japanese-to-English speech translation system: ATR-MATRIX. *ICSLP-98*, pages 2779-2782.
3. Jan Alexandersson, Norbert Reithinger and Elisabeth Maier. 1997. Insights into the Dialogue Processing of VERBMOBIL. In *Proceedings of ANLP*, pages 33-40.
4. Takahiro Ikeda, Shinichi Ando, Kenji Satoh, Akitoshi Okumura and Takao Watanabe. 2002. Automatic Interpretation System Integrating Free-style Sentence Translation and Parallel Text Based Translation. In *Proceedings of ACL Workshop on Speech-to-Speech Translation*, pages 85-92.
5. Ryosuke Isotani, Kiyoshi Yamabana, Shin-ichi Ando, Ken Hanazawa, Shin-ya Ishikawa, Tadashi Emori, Ken-ichi Iso, Hiroaki Hattori, Akitoshi Okumura, Takao Watanabe. 2002. An Automatic Speech Translation System on PDAs for Travel Conversation. In *Proceedings of ICMI-02*, pages 211-216.
6. Kiyoshi Yamabana, Shin-ichi Ando, Kiyomi Mimura. 2000. Lexicalized Tree Automata-based Grammars for Translating Conversational Texts. In *Proceedings of COLING 2000*, pages 926-932.
7. David S. Linthicum. 2000. Enterprise Application Integration. Addison-Wesley.